KnE Social Sciences

IC-HEDS 2019
International Conference on Humanities, Education, and Social Sciences
Volume 2020

Knowledge E
enriching | engaging | empowering

Conference Paper

The Estimates Item Parameter for Multidimensional Three-Parameter Logistics

Ode Zulaeha[1], Wardani Rahayu[2], and Yuliatri Sastrawijaya[3]

[1]Education Research and Evaluation, State University of Jakarta (UNJ), Jakarta, Indonesia
[2]Education Research and Evaluation, State University of Jakarta (UNJ), Jakarta, Indonesia
[3]Faculty Technique of Electrical, State University of Jakarta (UNJ), Jakarta, Indonesia

Abstract

The purpose of this study is to measure the accuracy of item parameters and abilities by using the Multidimensional Three-Parameter Logistics (M3PL) model. M3PL is a series of tests that measure more than one dimension of ability (θ). Item parameter estimation and the ability to model M3PL are reviewed based on a sample size of 1000 and test lengths of 15, 25, and 40. Parameter estimations are obtained using the Wingen software that is converted to BILOG. The results show that the estimate obtained with a test length of 15 displays a median correlation of 0.787 (high). The study therefore concludes that the level of difficulty of the questions is higher or the questions given to respondents are more difficult, so many respondents guessed the answers. The results of the estimated grain parameters and capabilities indicated that scoring based on sample size greatly affects the stability of the test length. By using the M3PL model, parameters can be measured pseudo-guessing, parameters b and parameters a. MIRT is able to explain interactions between the items on the test and the answers of the participants. The estimated results of the item parameters and the ability parameters of the participants also proved to be accurate and efficient.

Keywords: Multidimensional Three-Parameter Logistics (M3PL), distribution parameter, test length

Corresponding Author:
Ode Zulaeha
zulaehapepunj@gmail.com

Published: 11 November 2020

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the IC-HEDS 2019 Conference Committee.

1. Introduction

To measure a person's ability in a particular field, tests are usually carried out using test instruments so that his competence can be known. The test is one of the measurement tools most often used in the fields of education and psychology. In practice, the test should be objective, transparant, accountable, and non-discriminatory. A test kit should only be unidimensional, meaning that each test item only measures one ability. Assumptions can only be demonstrated if the test contains one factor that measures the achievement of a subject. The purpose of using the test is very diverse, but always with regard to one thing, namely the use of test scores to make a decision. the purpose

of the test is to test whether theoretical model on how to use test scores for certain purposes which so far may have been used is often quite reliable and trustworthy [1].

To get high-quality instruments, it is necessary to estimate the test items and capabilities. For this reason, it is necessary to estimate the items and capabilities. Various techniques can be used such as classical theory or modern theory. Along with the development of science in the field of psychometrics today modern theories are in great demand, such as the item response theory. The item response theory model produces item parameters that are independent of the test takers and test participant parameters that are independent of the set of items tested [2]. The invariant nature of the item response theory model makes theoretical item response theory can be used to problems solve that cannot be solved by classical test theory. In item response theory, there are assumptions that must be fulfilled, namely local independence and unidimension [3].

Local independence in item response theory occurs if the ability to influence the performance of a test is constant, meaning that the test participant's response in answering a test item is statistically independent of the test participant's response in answering other items. In other words, the assumption of local independence is stated that there is no correlation between test takers' responses to different items. This means that the ability expressed in the model is the only factor influencing test takers' responses to items. One of the assumptions that must be fulfilled in item response theory (IRT) is unidimension. Unidimensional means that the test only measures one certain ability. Unidimensional assumptions, in some cases as a whole point to measure the same ability [4]. Argue that what happens is that many tests measure more than one ability (multidimensional).

The MIRT (Multidimensional Theory Response Item) can be in the form of dichotomous or politomic score items. This study uses dichotomous data. To compile the matrix in the compiled MIRT expressed in the i-th row and j-th column elements. The items are stated in i (i = 1,..., n) and participants are stated in j (j = 1,..., N) [5]. The MIRT there are two models, namely compensatory and noncompensatory. According to the compensatory model allows high abilities in one dimension to obtain compensation in low abilities in other dimensions in relation to the probability of answering correctly [6]. Conversely, the noncompensatory model does not allow high capability on one to get compensation on low ability on other dimensions. For the compensatory model in the case of three-dimensional items, a test participant with very low ability in one dimension and very high ability in another dimension can answer the test item correctly. There are two types of compensatory models, namely the logistical MIRT model and the normal active model

of [7] by expressing a linear combination of multidimensional capabilities in rank on the probability formula to answer correctly. In this linear model, the low one or more capabilities can be compensated in other dimensions. Because compensation is a linear combination characteristic. The model used in this study is a multidimensional model of 3 logistical parameters (M3PL). Uses a three-parameter dichotomy logistic model into a multidimensional context, this M3PL model was developed by [8] using the following generalizations:

$$P(Xij = 1|j, ai, ci, di) = i + (1 - i)1 + \exp(.+)$$

Where, P (Xij = 1| j, ai, ci, di) is the probability of student j's correct response to item i; j is vector of student j's ability; ai is vector of item i slope; ci ∈ (0,1) is guessing parameter; and di is intercept parameter. Vector ai, j have the same elements m, which is the number of dimensions.

The M3PL model was designed to account for observed empirical data such as that provided [9] shows that examinees with low capabilities still have a probability of correct response. As a result, this model contains a single lower asymptote or guessing parameter ci ∈ (0,1) to specify such probability for examinees with very low value in θ. Theoretically, the interval of ci is between 0 and 1. In reality since ci ≥ 0,35 is often omitted from the test bank [10], ci ranges from 0 to 0,35.

This study discusses three logistical parameters of MIRT. The three-parameter model namely, discriminant, difficulty, and pseudo-gessing based on test lengths of 15, 25, and 40 with 1000 sample sizes.

2. Methods

The purpose of this study was to measure the accuracy of grain parameters and capabilities by using a model Multidimensional 3 Parameter Logistics (M3PL) reviewed based on 1000 sample sizes and test lengths of 15, 25, and 40. This study uses the UN SMP package data from the Education Assessment Center (PUSPENDIK) for DKI Jakarta in 2015 about mathematics. The dependent variable in this study is the RMSE of the correlation values generated using a model M3PL. As for the independent variables, namely parameters a, b, and c. The method used was a quantitative method. The estimated grain parameters and capability parameters are obtained by using the Wingen software combine to BILOG. Analysis of estimation results is done by comparing RMSE and correlation resulting from.

3. Results and Discussion

The following table presents the results of parameter analysis items a, b, and c using dichotomous data with a multidimensional model of 3 logistical parameters with a test length of 40.

The results of the study can be described in each table. In the table shows that from the simulation results with a test length of 40 items with multidimensional characteristics of 3 logistic parameters using a dichotomous scale for parameter discriminant (b), the average value is 1, the average parameter difficulty (a) is -1, and the average parameter pseudo-gessing (c) is value of 0. This indicates that the parameter a is higher or very influential to determine the parameters of the test item. This means that the value of difficulty is higher than the value of discriminant and pseudo-gessing. This shows that the level of difficulty index is greater than the calculation result, so the questions are too easy so that the probability of correct answers of test takers is greater.

The following will describe the results of the parameter estimation based on the length of the test. The description is based on the length of the test 15 in table 2.

Based on generated response data using WINGEN, and parameter estimation using BILOG software for 1000 sample sizes and 15 test lengths replicated 10 times. The results show that the estimate obtained with a test length of 15 shows a median correlation of 0.787 (high). This shows that the use of a sample size of 1000 with a test length of 15 items is very good based on the correlation value where the average is 0.7. This means that the estimated results of the item parameters by using the length of the test a little more influence on the ability of participants. For the sample size of 1000 and the length of the 25 tests that were replicated 10 times, the correlation value of the median was 0.664 or moderate. This means that the use of a sample size of 1000 with a test length of 25 is not very good, this can be seen from the average correlation of 0.6. Similar to using a test length of 25 items. For the sample size of 1000 and the length of the test 40 items obtained a correlation value of a median of 0.664 or medium category means that the estimated parameter parameters do not really affect the ability of participants.

From table 3 above, it shows that each item parameter with different test lengths shows different values. From the table it can be seen from the difficulty item (b) is higher with a 15 item test length of 0.906, for a 25 item test length of 0.867, and for a 40 item test length of 0.96. This means that the level of difficulty of the test items is very large. Questions given to participants are more difficult. While the pseudo-guessing parameter (c) for a test length of 15 items was 0.893, the length of the test was 25 items by 0.957, and the length of the test by 40 items was 0.97. This means that the guessing

Table 1: Estimated Item Parameters With M3PL

Test Lenght	M3PL Item Characteristics	Item Parameter a	b	c
1	3PLM 2	1,209	-1,722	0,055
2	3PLM 2	1,754	0,591	0,081
3	3PLM 2	1,728	-1,343	0,054
4	3PLM 2	1,947	-1,194	0,068
5	3PLM 2	0,736	0,409	0,002
6	3PLM 2	0,409	-1,610	0,089
7	3PLM 2	0,499	1,381	0,100
8	3PLM 2	1,272	-0,896	0,003
9	3PLM 2	1,630	-0,101	0,085
10	3PLM 2	1,038	1,534	0,022
11	3PLM 2	0,724	-0,697	0,076
12	3PLM 2	1,486	-1,952	0,009
13	3PLM 2	0,554	-0,633	0,005
14	3PLM 2	1,119	-0,410	0,081
15	3PLM 2	0,870	-1,261	0,064
16	3PLM 2	1,782	1,923	0,036
17	3PLM 2	0,659	1,525	0,068
18	3PLM 2	1,915	-0,048	0,096
19	3PLM 2	1,505	-1,870	0,005
20	3PLM 2	1,096	-0,249	0,030
21	3PLM 2	0,511	-0, 502	0,048
22	3PLM 2	0,896	-1,087	0,099
23	3PLM 2	1,623	0,861	0,021
24	3PLM 2	1,936	-1,141	0,029
25	3PLM 2	1,013	0,133	0,068
26	3PLM 2	0,563	1,721	0,028
27	3PLM 2	1,769	1,069	0,056
28	3PLM 2	1,806	1,433	0,072
29	3PLM 2	1,535	0,312	0,008
30	3PLM 2	1,676	-0,356	0,052
31	3PLM 2	1,166	-1,542	0,054
32	3PLM 2	1,030	-1,912	0,041
33	3PLM 2	1,234	0,264	0,040
34	3PLM 2	1,680	0,031	0,025
35	3PLM 2	1,639	-1,887	0,045
36	3PLM 2	1,376	-1,576	0,022
37	3PLM 2	0,781	1,880	0,079
38	3PLM 2	1,524	1,378	0,059
39	3PLM 2	0,736	-0,692	0,073
40	3PLM 2	1,883	-1,405	0,098

Table 2: Results of Estimated Item Parameters With Test Length 15, 25, and 40

Replikasi	Proficiency Parameter Report (N=15)		Proficiency Parameter Report (N=25)		Proficiency Parameter Report (N=40)	
	Corr.	RMSE	Corr.	RMSE	Corr.	RMSE
1	0,794	933657,5	0,655	992214,6	0,629	1022288
2	0,785	929053,4	0,663	996345,8	0,646	1026089
3	0,788	937981,4	0,669	995188,8	0,635	1030588
4	0,786	938160,7	0,663	975822,5	0,637	1025104
5	0,79	926342,9	0,671	982328,3	0,646	1033348
6	0,792	934581,9	0,666	990173,9	0,644	1025356
7	0,774	929346,7	0,661	983413,7	0,633	1030554
8	0,776	934752,4	0,667	988648,9	0,633	1031710
9	0,788	931338,2	0,664	992608	0,643	1032769
10	0,783	935659,3	0,656	995289,4	0,634	1024116
	Median (corr) = 0,787		**Median (corr) = 0,664**		**Median (corr) = 0,636**	

Table 3: Estimates Item Parameter a, b, and c Based on Test Lengths 15, 25 and 40.

Parameter	Test Lenght		
	15	25	40
a	0,237	0,213	0,05
b	0,906	0,867	0,96
c	0,893	0,957	0,97

answer given is very good so the participants are confused to determine the correct answer. While the distinguishing power of item (a) is very lacking. This can be shown from the estimation results which are made on average reaching 0.2 for test lengths 15 and 25. While for test length 40 is 0.0.

This research was influenced by the number of sample sizes with the length of the test used. There are three types of test lengths used. The simulation results using the Wingen software combined with BILOG which were replicated 10 times showed that the estimated grain parameters using M3PL were very good. This can be seen from the results of data simulations that the quality of items is in accordance with the ability of participants. From the results of this study indicate that a good test length is used for the number of samples of 1000, namely as many as 15 item test lengths.

This research was conducted based on the results of a review of previous articles about the nonlinear model or better known as the multidimensional model. In

classical theory three parameters are measured namely, difficulty, discriminant, and gessing. Likewise in modern theory or known as item response theory (IRT). In the IRT multidimensional model, three aspects are measured namely, discriminant (slope), difficulty (threshold), and pseudo-gessing and other abilities that can be measured on multidimensional items. There is an assumption on IRT that is unidimensional, where the test only measures one ability. However, Folk and Green believes that what happens is that many tests measure more than one ability (multidimensional) [11]. Multidimensional is a test model that distinguishes the ability of participants, where the test is easy for participants who have high ability and low test/difficulty for participants who have low ability [12]. In studies that carry out simulation tests, revealing the effect of sample size and test length on the stability of the estimated grain parameters and the ability of test takers on IRT is seen based on one parameter, two parameters, and three parameters [13].

Based on the results of research Nandakumar, 1994; Richard A. Ashley and Douglas M. Patterson, 1989; Daniel J. Bauer, 2005; Groves, Kevin S Vance and Charles M, 2015; Rosemary A. Abbott, Caroline Skirrow, Martha Jokisch, Maarten Timmers, Johannes Streffer, Luc van Nueten, Michael Krams, Angela Winkler, Noreen Pundt, Pradeep J. Nathan, Philippa Rock, Francesca K. Cormack, Christian Weimar, 2019 show that Usage Nonlinear models have an average value with high correlation. Multidimensional research based on unidimensional assumptions will experience obstacles in assessing multidimensional tests (Ackerman, 1989; Cheng, Wang, and Ho, 2009; DeMars, 2006; Dirir and Sinclair, 1996; Oshima and Miller, 1990; Reise, 1990 Moore Moore, and Haviland, 2010; Yao, 2011). Multidimensional tests will experience inaccuracies if assessed according to the unidimensional paradigm. To correctly answer test items, test participants often need more than one latent attribute so it is called multidimensional [14]. So it is very important to conduct multidimensional research, specifically using three logistical parameters.

4. Conclution

Based on the results and discussion, it shows that the item parameter estimation is more effective if it uses a small test length and in compiling the item needs to be considered the distinguishing power of an item in order to see the difference in the probability of the response answering the item correctly. So, the M3PL model is very effective for estimating item parameters and response capabilities.

References

[1] Husen, U. (2011). *Research Methods for Thesis and Business. (2nd ed.)*. Jakarta: PT Raja Grafindo Persada.

[2] Xitao, F. (1998). Book Review of Structural Equation Modeling With LISREL, PRELIS, and SIMPLIS: Basic Consept Applications, and Programming by B.M Byrne. *Educational and Psychological Measurement*.

[3] Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. CA: Sage Publication Inc.

[4] Yanyan, S. and Wikle, C. K. (2007). Comparing Multidimensional and Unidimensional Item Respon Theory Models. *Educational and Phsychological Measurement*, http://doi.org/10.1177/0013164406296977.

[5] Reckase, M. D. (1997). The Past and Future of Multidimensional Item Respon Theory. *Applied Psychological Measurement*, **doi:10.1177/0146621697211002**.

[6] Reckase, M. D. (1985). The Difficulty of Test Items That Measure More Than One Ability. *Applied Psychological Measurement*.

[7] Samejima, F. (1974). Normal Ogive Model on the Continuous Response Level in the Multidimensional Space. *Psychometrika*.

[8] Reckase, M. D. (1996). A Linear Logistic Multidimensional Model. In W. J. van der Linder and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*. New York: Springer-Verlag, pp. 271–286.

[9] Lord, F. M. (1980). A*pplication of Item Response Theory to Practical Testing Problems*. Hillsdale: Lawrence Erlbaum Associates.

[10] Baker, F. B. (2001). *The Basic of Item Response Theory*. USA: ERIC Clearinghouse on Assessment and Evaluation.

[11] Folk, V. G. and Green, B. F. (1989). Adaptive Estimation when the Unidimensionality Assumption of IRT is Violated. *Applied Psychological Measurement*.

[12] Yalcin, I. (1995). *Nonlinear Factor Analysis*. *Retrospective Theses and Dissertations*. USA: IOWA State University.

[13] McDonald, R. P. (1997). Multidimensional Normal Ogive Model. In W. J. Van der Linden and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*. New York: Springer-Verlag, pp. 257-269.

[14] Hambleton, R. K., and Cook, L. L. (1977). Latent Trait Models and their Use in the Analysis of Educational Test Data. *Journal of Educational Measurement*.

[15] Ackerman, T. A. (1994). Using Multidimensional Item Response Theory to Understand What Items and Tests are Measuring. *Applied Measurement in Education*.